NOI TEHNICI UTILIZATE ÎN ANALIZA AUTOMATĂ A TEXTULUI

Mihai ISTRATE, *Universitatea "Constantin Brancuşi" din Tg-Jiu*

***REZUMAT**: analiza automată a textelor în limbaj natural este una dintre cele mai importante sarcini descoperirea de cunoştinţe pentru orice organizaţie. Conform Gartner Group, aproape 90% din cunoştinţele disponibile la o organizaţie de azi sunt dispersate în mormane întregi de documente îngropate într-un text nestructurat. Analiza unor volume uriaşe de informaţii de tip text conduce adesea la luarea unor decizii informate şi corecte de afaceri. Metodele tradiţionale de analiză bazate pe statistici nu ajuta la procesarea textelor nestructurate şi societatea este în căutare de noi tehnologii pentru analiza de text. Există o varietate de abordări pentru a analiza textelor limbajului natural, dar cele mai multe dintre ele nu oferă rezultate care ar putea fi aplicate cu succes în practică. Acest articol se concentrează pe idei recente şi implementări practice în acest domeniu.*

***CUVINTE CHEIE:** text mining, analiza textului, reţele neuronale*

1.INTRODUCERE

Luarea deciziilor corecte necesită deseori analizarea volumelor mari de informaţii textuale. Cercetătorii, analiştii, editorii de reviste, societăţile cu capital de risc, avocaţii, specialiştii, chiar şi studenţii se confruntă cu diferite sarcini ce implica analiza de text [1].

Grămezi imense de informaţii se acumulează în arhivele text ale numeroaselor agenţii de ştiri, biblioteci, corporaţii, PC-uri individuale, şi reţelei Web. Cantitatea de informaţii stocate proliferează într-un ritm dezastruos, iar ochii şi creierul uman sunt puşi din ce în ce mai mult în imposibilitatea de a face faţă provocărilor acestei creşteri. Omenirea este în căutarea de asistenţă electronică inteligentă care să ajute în proiectele de analiză de text.

Prin urmare, o noua tehnică a fost necesară pentru a rezolva acest tip de probleme. Şi aşa

NEW TECHNIQUES USED IN AUTOMATED TEXT ANALYSIS

M. Istrate, *„Constantin Brancusi" University of Tg-Jiu*

ABSTRACT:

Automated analysis of natural language texts is one of the most important knowledge discovery tasks for any organization. According to Gartner Group, almost 90% of knowledge available at an organization today is dispersed throughout piles of documents buried within unstructured text. Analyzing huge volumes of textual information is often involved in making informed and correct business decisions. Traditional analysis methods based on statistics fail to help processing unstructured texts and the society is in search of new technologies for text analysis. There exist a variety of approaches to the analysis of natural language texts, but most of them do not provide results that could be successfully applied in practice. This article concentrates on recent ideas and practical implementations in this area.

***KEY WORDS**: text mining, text analysis, neural network.*

1. INTRODUCTION

Making correct decisions often requires analyzing large volumes of textual information. Researchers, analysts, magazine editors, venture capitalists, lawyers, help desk specialists, and even students are faced by various text analysis tasks [1].

Huge piles of information accumulate in numerous text repositories held at news agencies, libraries, corporations, individual PCs, and the Web. The amount of stored information proliferates at a disastrous rate, and the human eyes and brain are increasingly unable to meet the challenges of this growth. Mankind is searching for intelligent electronic assistants to help with text analysis projects.

Therefore a new technique was necessary to resolve this kind of problems. And so was born Text Mining. It derives from Data Mining and is getting more and more

s-a născut Text Mining. Aceasta derivă din Data Mining şi capătă din ce în ce mai mult atenţia în cadrul organizaţiilor de afaceri de azi. Rolul său este de analiză cu privire la volume mari de text şi de a găsi modele ascunse prin mijloace automate sau semiautomate.

Text Mining este un proces de extragere a noilor cunoştinţe dispersate în întregime în documente text şi utilizează aceste cunoştinţe pentru a organiza mai bine informaţiile pentru o viitoare folosire a lor.

2.APLICATII SI OBIECTIVE ALE ANALIZEI TEXTULUI

În primul rând, ar fi de dorit să fie în măsură să distileze automat într-o formă concisă sensul unui text şi să stocheze rezultatele ca o listă a celor mai importante concepte din text printr-un hyperlink cu locurile corespunzătoare din textul original. Această procedură ar putea oferi un nou mecanism eficient de navigare prin texte, crearea automată de rezumate ale documentelor, gruparea şi clasificarea textelor, comparaţia de texte, precum şi extragerea de informaţii din limbajul natural. Atingerea acestei funcţionalităţi ar putea avea profunde implicaţii practice pentru activităţile noastre de zi cu zi în ceea ce priveşte procesarea textelor.

În marea majoritate, noi toţi suntem nevoiţi să ne ocupăm cu revizuirea unor volume mari de informaţii textuale. În acelaşi timp, pentru unele profesii capacităţile de analiză automată inteligentă de text pot fi critice. O funcţie de rezumare automată a textului ar putea fi utilizată de către guvern şi analiştii din mediul de afaceri, editorii de reviste, societăţile cu capital de risc, avocaţi, şi studenţi, care doresc să poată vedea rezumatele exacte înainte de a se cufunda în documentul complet. O navigare eficientă printr-o bază de text, precum şi rezumare, grupare şi clasificare a textelor, ar putea spori eficienţa de lucru cu baze de text mari, cum ar fi documentele academice (pentru cercetători), fluxul de ştiri electronice (pentru

attention in today's business organizations. Its role is about analyzing large volumes of text and finding hidden patterns using automatic or semiautomatic means.

Text Mining is a process of extracting new knowledge dispersed throughout text document and utilizing this knowledge to better organize information for future reference.

2. TEXT ANALYSIS TASKS AND APPLICATIONS

First, one would like to be able to automatically distill the meaning of a text in a concise form and store the results as a list of the most important concepts from the text hyperlinked with the corresponding places in the original text. This procedure would provide a new efficient mechanism for navigation through texts, automated creation of summaries of documents, clustering and classification of texts, comparison of texts, as well as natural language information retrieval.

Achieving this functionality could have profound practical implications for our everyday text processing activities.

By and large, we all have to deal with reviewing large volumes of textual information. At the same time, for some professions automated intelligent text analysis capabilities can be critical. An automated text summarization function could be used by government and business analysts, magazine editors, venture capitalists, lawyers, and students, who wish to see accurate summaries before plunging into the full documents. An efficient navigation through a textbase, as well as summarization, clustering and classification of texts, could enhance the effectiveness of working with large textbases including academic documents (for researchers), electronic news flow (for marketers and investment bankers), and e-mail systems (for all users). An automated

marketing si bancheri) şi sistemele de e-mail (pentru toţi utilizatorii). O clasificare automată a mesajelor de intrare către diferite grupuri de subiect şi priorităţi prin analiza conţinutului lor, precum şi recuperarea eficientă a acestora la o dată ulterioară ar putea ajuta la vindecarea traumei din experienţa noastră de e-mail. Capacitatea de extragere de informaţii semantice ar putea salva milioane de ore/om, prin creşterea relevanţei şi preciziei de căutare într-o baze de date sau de navigare pe Internet. Grupând o colecţie de documente care reprezintă reacţia presei la mişcările recente de marketing ale companiei dumneavoastră şi ale concurenţilor dumneavoastră ar putea ajuta la evaluarea eficienţei campaniei dumneavoastră de marketing. O combinaţie a tuturor acestor funcţii cu o capacitate de extragere a informaţiilor din limbaj natural ar putea facilita crearea unei noi generaţii de puternice şi inteligente soluţii de Help Desk Support şi Call Center.

Perspectivele se arată înfloritoare, dar problema este că toate încercările de a construi sisteme practice pentru analiza automată a textelor limbajului natural nu au produs rezultate satisfăcătoare până în prezent. Sistemele create, de obicei, funcţionează bine doar într-o aplicaţie dintr-un anumit domeniu şi necesită o intervenţie umană semnificativă şi costisitoare în cazul de adaptare a sistemului la un domeniu nou. Astfel, obiectivul ar fi să fie dezvoltată o nouă abordare pentru o analiză mai versatilă şi mai automată de texte din diferite discipline. Să discutăm pe scurt tehnicile tradiţionale de analiză de text, în scopul de a identifica părţile lor puternice şi părţile lor slabe.

3.TEHNICI DE ANALIZA A TEXTULUI

3.1. Abordarea tradiţională a analizei textului

În general, sistemele bazate pe abordările tradiţionale analizează un text scris în limbaj natural într-un anumit fel, la nivelul fiecărei

classification of incoming messages to different subject groups and priorities through the analysis of their contents, as well as their efficient retrieval at a later time could help to heal the trauma of our email experience. The semantic information retrieval capability could save millions of man-hours by increasing the relevance and precision of a database search or Internet surfing. Clustering a collection of documents that represent the press reaction to the latest marketing moves of your company and your competitors could help assess the effectiveness of your marketing campaign. A combination of all of these functions with a natural language information retrieval capability could facilitate creating a new generation of powerful and intelligent corporate Help Desk and Call Support Center solutions.

The prospects look bright, but the problem is that all the attempts to build practical systems for automated analysis of natural language texts have not produced satisfactory results thus far. The created systems usually work well only in a certain application field and require significant and costly human interference at the stage of tuning the system to a new field. Thus the objective would be to develop a new approach for more versatile and automated analysis of texts from different subjects. Let us first briefly discuss traditional text analysis techniques in order to identify their strong and weak sides.

3. TEXT ANALYSIS TECHNIQUES

3.1. Traditional approach text analysis

In general, systems based on traditional approaches analyzed a natural language text in a certain way at the level of individual sentences [2].

The objective was to create a semantic representation of a sentence in the form of structured relations between important words comprising this sentence. To solve this task, various predeveloped linguistic molds were

propoziţii [2]. Obiectivul era de a crea o reprezentare semantică a unei propoziţii sub formă de relaţii structurate intre cuvinte importante care formează această propoziţie. Pentru a rezolva această sarcină, diverse forme lingvistice au încercat analiza propoziţiei şi a componentelor sale. Când un tipar se potrivea bine propoziţiei, o construcţie corespunzătoare semantică era asociata cu propoziţia. Această tehnică a oferit o prima orientare buna pentru a înţelege sensul unui text. Dar, aşa cum se dovedeşte, principala problemă a acestei abordări este ca pot exista prea multe tipare diferite necesare ce trebuie construite pentru a analiza diferite tipuri de propoziţii. În plus, lista de construcţii excepţionale în această abordare creste rapid in forme exagerat de mari. Cu alte cuvinte, această abordare funcţionează bine numai pentru o submulţime limitata de texte in limbajului natural.

Una dintre ramurile tradiţionale ale Inteligentei Artificiale (AI), cunoscuta ca domeniu de comunicare prin intermediul limbajului natural computer-uman, este in principal axata pe procesarea automată a textelor. Aceasta ramura include traducere automată, căutare semantica de informaţii, şi crearea de sisteme expert. Aici sunt puse în aplicare metode pur lingvistice pentru analiza semantica a textului. Rezultatele analizei sunt reprezentate sub forma unei reţele semantice care afişează o listă cu cele mai importante cuvinte din text şi relaţiile dintre ele. O reţea semantică este o reprezentare text convenabila adesea folosita în ştiinţele cognitive. Trebuie remarcat faptul că un set de reguli lingvistice odată creat funcţionează foarte bine doar cu texte care au acelaşi subiect pentru care aceste reguli au fost dezvoltate. Astfel, analiza efectuată este puternic dependentă de cunoştinţele generale din domeniul analizat. Acest lucru implică, de asemenea că un expert uman trebuie să fie implicat în stadiul de dezvoltare a normelor lingvistice pentru un anumit subiect. O astfel de abordare funcţionează bine in crearea de sisteme expert care sunt utilizate numai într-un câmp de aplicaţie unică. Cu toate acestea,

tried with the sentence and its components. When a mold matched the sentence well, a corresponding semantic construction was associated with the sentence. This technique provides a good first guidance for understanding the meaning of a text. But as it turns out, the main problem with this approach is that there can be too many different molds that one needs to build for analyzing different types of sentences. In addition, the list of exceptional constructions in this approach quickly grows prohibitively large. In other words, this approach works well only for a limited subset of natural language texts.

One of the traditional branches of Artificial Intelligence (AI), known as the field of natural language computer-human communication, is mainly devoted to automated processing of texts. This branch includes machine translation, semantic search for information, and creation of expert systems. Here purely linguistic methods are implemented for the analysis of the text semantics. The results of the analysis are represented in the form of a *semantic network* displaying a list of the most important words from the text and relations between them. A semantic network is a convenient text representation object often used in cognitive sciences. It should be noted that a set of the created linguistic rules works well only with texts within the subject for which these rules have been developed. Thus the performed analysis is strongly dependent on the background knowledge of the analyzed field. This also implies that a human expert must be involved at the stage of the development of linguistic rules for a subject. Such an approach works well for the creation of expert systems that are utilized only in a single application field. Yet, in order to successfully analyze texts from arbitrary fields, one needs to employ more general algorithms.

Another approach applicable to processing unstructured texts, artificial Neural Networks

în scopul de a analiza cu succes texte din diverse domenii, trebuie dezvoltați algoritmi mai generali.

O altă abordare care se aplică la prelucrarea textelor nestructurate, rețelele Neuronale Artificiale (NN), a fost dezvoltata cu speranța că un mediu prelucrare artificial omogen realizat în mod similar cu conexiunile creierului uman ar putea procesa informațiile, într-adevăr, întocmai ca si creierul uman. Din nou, s-a demonstrat că sistemele bazate pe această abordare sunt capabili să rezolve cu succes sarcini simple de analiză. Cu toate acestea, în general, un suport omogen de prelucrare nu este prea potrivit pentru analiza de informații lingvistice structurate. Dezvoltarea unui nou tip de procesare media structurate este, aşadar, necesară pentru a putea face față acestei sarcini.

Rezumând, ambele abordări AI şi NN oferă perspective importante în aceasta problemă, dar au demonstreze doar un succes limitat în aplicații practice. De fapt, tehnicile cele mai promițătoare pentru analiza textelor în limbaj natural rezulta din suprapunerea celor două domenii. O abordare foarte interesanta o reprezintă folosirea unui mediu pentru analiza de text format din unități de procesare paralelă, ca şi în abordarea NN, în timp ce structurarea acestui mediu sa fie în funcție de modele cognitive de AI, în cazul în care sarcina este divizată într-un număr de subprobleme conectate prin fluxuri de informații reprezentate în condițiile modelelor cognitive. În acest fel, poate fi formata o construcție mai bogata şi mai complexa adecvata pentru analiza de texte.

Informațiile sunt prelucrate automat ca in NN, dând naştere, în acelaşi timp la structuri semantice, care sunt adesea întâlnite în AI.

3.2. O nouă abordare a analizei textelor

În noua metodă hibrid, textul este considerat ca o secvență de simboluri organizate în cuvinte şi propoziții. Această secvență este mutată printr-o fereastră de lungime variabilă (de le doua pana la (NN), was developed with the hope that a homogeneous artificial processing media made out of connected elements similar to the brain neurons could indeed process information similarly to the human brain. Again, it has been demonstrated that systems based on this approach are capable of successfully solving simple analysis tasks. However in general, a homogeneous processing media is not suited well for the analysis of linguistically structured information. Developing a new type of structured processing media is required for tackling this task.

Summarizing, both AI and NN approaches provide important insights into the problem but demonstrate only limited success in practical applications. In fact, the most promising techniques for the analysis of natural language texts reside in the overlap of the two fields. One very interesting approach is to employ for text analysis the media consisting of parallel processing units, as in the NN approach, while structuring this media according to cognitive models of AI, where the task is split into a number of subtasks connected by information flows represented in terms of cognitive models. In this way, a more rich and complex construction suitable for the further text analysis can be formed.

The information is processed automatically as in NN, while at the same time giving birth to semantic structures, which are often encountered in AI.

3.2. New approach to text analysis

In the new hybrid method, the text is considered as a sequence of symbols organized into words and sentences. This sequence is moved through a window of variable length (from two to twenty symbols can be seen simultaneously), shifting it by one symbol at a time. The snapshots of the text fragments visible through the window are

douăzeci de simboluri pot fi văzute simultan), schimbând câte un simbol, la un moment dat. Instantanee din fragmente de text vizibile prin fereastra sunt înregistrate în neuroni dinamic adăugaţi.

Reţeaua neuronală ierarhică creată conţine mai multe straturi: acele fragmente care apar în text mai mult de o dată sunt stocate în neuroni care aparţin la nivelurilor superioare ale reţelei. Această reţea neuronală realizează dicţionarele de frecvenţă pe baza mai multor nivele a elementelor de text diferite (litere, silabe, tulpini, morfeme, cuvinte, şi fraze). Cuvintele sunt selectate ca elemente operaţionale de bază, în timp ce alte elemente sunt folosite ca informaţii auxiliare în timpul analizei

recorded in dynamically added neurons.

The created hierarchical neural network contains several layers: those fragments that occur in text more than once are stored in neurons that belong to the higher levels of the network. This neural network realizes frequency-based multi- level dictionaries of different text elements (letters, syllables, stems, morphemes, words, and phrases). Words are selected as basic operational elements, while other elements are used as auxiliary information during the analysis

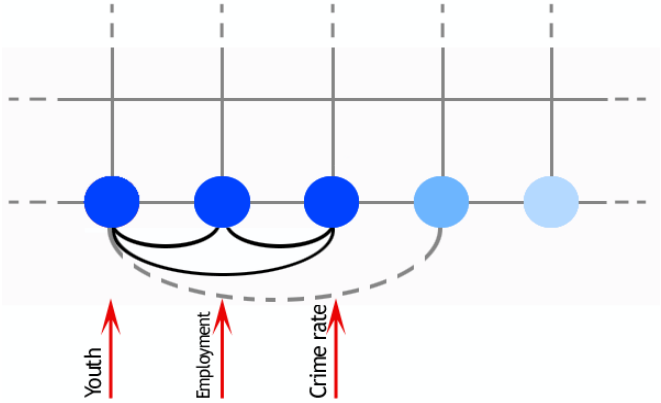

Figura 1. Creşterea dinamică a reţelei neuronale înregistrează noi fragmente de text
Figure 1. Dynamically growing neural network records new text fragments

În mod ideal, se doreşte a se scăpa de toate cuvintele suplimentare şi banale, care nu au sens semantic. De asemenea, s-ar dori identificarea tulpinilor de cuvânt, adică separarea de prefixe, sufixe, şi terminaţii (morfeme). Acest pas se numeşte *preprocesare.*

Toate lucrările suplimentare vor fi efectuate numai cu tulpini, îmbunătăţind astfel calitatea analizei. De exemplu, cuvintele "înseamnă" şi "semnificative" vor fi identificate de acest sistem ca având aceeaşi tulpină.

Ideally, one wishes to get rid of all supplementary and commonplace words, which carry no semantic meaning. Also one would like to identify stems of the words, while separating prefixes, suffixes, and endings (morphemes). This step is called *preprocessing*.

All further work can be carried out with stems only, thus improving the quality of the analysis. For example, the words “mean” and “meaningful” will be identified as having the same stem by this system.

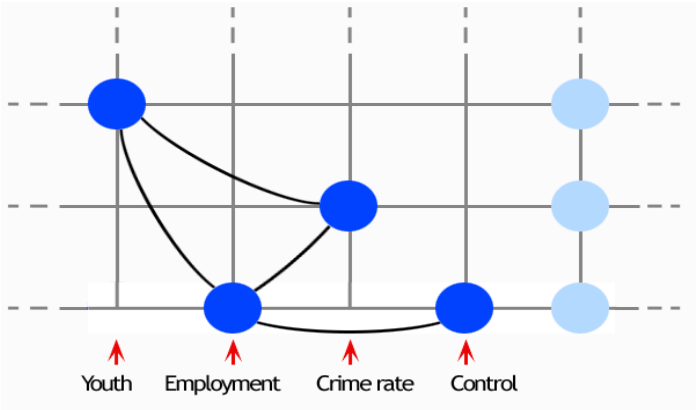

Figura 2.Reţeaua neuronală recurent ierarhică depistează frecvenţele şi relaţiile termenilor
Figure 2. Hierarchical recurrent neural network traces frequencies and relations of terms

De fapt, obţinerea unui mecanism eficient de preprocesare necesită reglaj fin al sistemului într-o limbă specificată în scopul de a filtra în mod eficient cuvinte suplimentare şi morfeme nativ pentru această limbă. S-ar putea utiliza aceeaşi reţea ierarhic neuronale pentru a construi un filtru pentru elementele nedorite. Atunci când prelucrează un corpus mare de texte din diverse subiecte, cuvinte suplimentare şi morfeme sunt fragmentele care apar cel mai frecvent în text. Prin colaborarea cu diverse fragmente de cuvinte, reţeaua ierarhică neuronale permite capturarea automată atât a cuvintelor suplimentare cât şi a morfemelor, în acelaşi timp. Reţineţi că această preprocesare este singurul loc unde dependenţa de limba intră în discuţie si tehnicile de noi analize şi unde unele îndrumări umane sunt de dorit. Toate celelalte componente ale acestei tehnologii sunt independente de limbă şi de a lucra la fel de bine cu texte, în orice limbă bazată pe un alfabet. Aplicând un prag reţelei neuronale dezvoltate pe astfel de texte, se creează un filtru care poate fi folosit mai târziu pentru separarea tulpinilor de cuvinte semantice importante pentru analize suplimentare. În timp ce efectuează analiza cu tulpini individuale, reţeaua deţine încă informaţii despre cuvintele complete.

Să presupunem că am reuşit să filtrăm elementele lipsite de sens şi să procesăm informaţiile semnificative. Nodurile reţelei neuronale dezvoltate deţin acum toate cuvintele importante şi combinaţii de cuvinte

In fact, obtaining an efficient preprocessing mechanism requires fine-tuning the system to a specified language in order to efficiently filter out supplementary words and morphemes native to this language. One could utilize the same hierarchical neural network in order to build a filter for unwanted elements. When processing a large corpus of texts from diverse subjects, supplementary words and morphemes are the fragments appearing most frequently in the text. By working with various fragments of words, the hierarchical neural network allows one to automatically catch both supplementary words and morphemes at the same time. Note that this preprocessing is the only place where language dependency enters in the discussion of the new analysis technique and where some human analyst guidance is desirable. All other components of this technology are language independent and work equally well with texts in any alphabet-based language. Applying a threshold to the neural network developed on such a corpus of texts, one creates a filter that can be used later for separating the stems of semantically important words for further analysis. While performing the analysis with individual stems, the network still holds the information about complete words.

Let us assume that we managed to filter out meaningless elements and process the significant information. The nodes of the developed neural network now hold all

din text, cu frecvenţele apariţiei lor. În acelaşi timp, aceeaşi reţea evaluează frecvenţa de apariţie în comun a diferitelor elemente semantice în cadrul unor unităţi structurale text, de exemplu propoziţii. Se obţine o structură de tip graf care conţine ponderile statistice ale cuvintelor în nodurile şi ponderile statistice ale aparițiilor în comun ale acestor cuvinte în legăturile dintre noduri. Acest grafic nu oferă încă o imagine semantică exactă a textului analizat. Mai trebuie încă să se adapteze ponderile individuale statistice ale cuvintelor şi a relaţiilor dintre ele pentru a oferi o reprezentare consistentă a textului. Ponderile de aceste cuvinte, care sunt strâns legate de alte cuvinte frecvente din text ar trebui să fie stimulate, şi vice-versa.

important words and word combinations from the text with the frequencies of their occurrence. Simultaneously, the same network assesses frequencies of joint occurrence of different semantic elements within certain structural text units, for example sentences. One obtains a graphlike structure that contains statistical weights of words in the nodes and statistical weights of joint occurrences of these words in the links.

This graph does not provide an accurate semantic picture of the analyzed text yet. One still needs to adjust individual statistical weights of the words and relations between them to provide a consistent text representation. The weights of those words, which are strongly related to other frequent words in the text should be boosted, and vice versa.

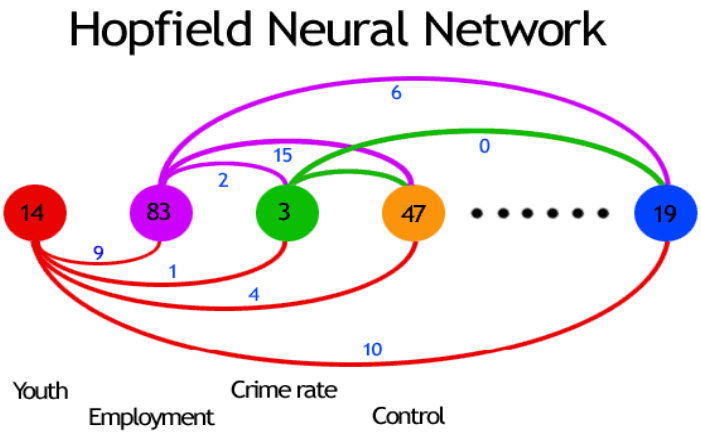

Figura 3. Reţeaua neuronală Hopfield detaliază acurateţea semantică a analizei
Figure 3. Hopfielfd-like neural network refines semantic accuracy of the analysis

Acest lucru este realizat prin atribuirea ponderile statistice de cuvinte individuale la nodurile dintr-o reţea unidimensională neuronale Hopfield, în care toţi neuronii sunt complet interconectate. În acelaşi timp, ponderile statistice ale relaţiilor dintre cuvinte sunt atribuite legăturile dintre nodurile individuale în această reţea. Când a fost lansată, această reţea Hopfield s-a remarcat prin modificarea ponderilor atribuite noduri şi legăturile între ele la o configuraţie stabilă corespunzătoare minimului unei funcţii de energie, ce caracterizează reţeaua. Ponderile

This is accomplished by assigning the statistical weights of individual words to the nodes in a one-dimensional Hopfield-like neural network where all neurons are completely interconnected. Simultaneously, the statistical weights of relations between words are assigned to the links between individual nodes in this network. When released, this Hopfield-like network evolves by changing the weights assigned to the nodes and links between them to a stable configuration corresponding to the minimum of an energy-like function characterizing the

renormate ale cuvintelor şi relaţiile dintre ele se numesc ponderi semantice şi structura reconfigurată de tip graf rezultată se numeşte o reţea semantică (care este o listă a celor mai importante cuvinte şi combinaţii de cuvinte din text şi a relaţiilor dintre ele).

Având în vedere că analiza unui text a fost realizată fără nici un recurs la orice cunoştinţe generale a subiectului de interes, sensul unui cuvânt în reţeaua semantică creată este definit doar de acele cuvinte, care sunt legate de acesta în reţea.

Corespunzător, cuvintele şi combinaţii de cuvinte care sunt cuprinse de o reţea semantică au un nume special - concepte semantice. Reţeaua semantică reprezintă un punct de vedere lingvistic exact şi concis a textului analizat. Această construcţie poate sta la baza tehnicilor multor analize suplimentare puse în aplicare de utilizator necesare funcţionalităţii de procesare a textelor.

4. SOFTWARE

În acest moment multe aplicaţii software sunt disponibile pentru analiza de text si multe altele sunt în curs de dezvoltare. Un exemplu de software prezentate mai jos, care utilizează tehnici de text mining este TextAnalyst [3], [4].

TextAnalyst este un instrument software unic pentru analiză semantică, navigare şi căutare a textelor nestructurate. O sinergie de tehnologii şi reţele lingvistice unice puse în aplicare de TextAnalyst asigură o viteză mare si precizie deosebită în analiza textelor nestructurate.

Software-ul TextAnalyst se remarcă prin:

- **Distilarea sensului unui text** - formarea şi exportarea cu acurateţe a unei reţele semantice a textului sau bazei de text. Această reţea reprezintă în mod concis sensul unui text şi serveşte drept bază pentru toate o analiză suplimentară.
- **Rezumatul exact al textului** - calitatea rezumatului este furnizată printr-o combinaţie echilibrată de metode lingvistice şi metode de investigare a

network. The renormalized weights of words and relations between them are called semantic weights and the resulting reshaped graph- like structure is called a semantic network (which is a list of the most important words and word combinations from the text and relations between them).

Since the analysis of a text has been performed with no recourse to any background knowledge of the subject of interest, the meaning of a word in the created semantic network is defined purely by those other words, which are related to it in the network.

Correspondingly, the words and word combinations comprising a semantic network have a special name - semantic concepts. The semantic network represents a linguistically accurate and concise picture of the analyzed text. This construction can lie in the foundation of many further analysis techniques implementing user- needed text processing functionality

4. SOFTWARE

At this time many software applications are available for text analysis and many others are in course of development. An example of software presented below which use text mining techniques is TextAnalyst [3],[4].

TextAnalyst is a unique software tool for semantic analysis, navigation, and search of unstructured texts. A synergy of unique linguistic and neural network technologies implemented in TextAnalyst ensures high speed and accuracy in the analysis of unstructured texts.

- **Distilling the meaning of a text** - formation and export of an accurate Semantic Network of the text or textbase. This network concisely represents the meaning of a text and serves as a basis for all further analysis.
- **Accurate summarization of texts** - the quality of the summary is provided by a balanced combination of linguistic and neural network investigation methods.

rețelei neuronale. Dimensiunea rezumatului este controlat prin intermediul pragului ponderilor semantice.

- **Explorarea textului concentrată pe subiect** - dicționarele specificate de utilizator de cuvinte excluse şi incluse permite anchetei să se concentreze pe un subiect ales.
- **Navigarea eficientă printr-o bază de text** - în baza de cunoştințe se poate naviga cu legături de la conceptele din rețeaua semantică până la propoziții în documentele care conțin o combinație considerată de concepte. Fraze individuale sunt la rândul lor, hyperlink-uri la acele locuri din textele originale unde au fost întâlnite.
- **Explicarea structurii temei textului** - o structură de arbore cu tema reprezentând semantica textelor investigate este creată în mod automat. Cele mai importante subiecte sunt amplasate mai aproape de rădăcina arborelui.
- **Gruparea textelor** - link-uri întrerupte reprezentând relațiile slabe din rețeaua semantică originală permit gruparea bazei de text.
- **Extragerea informațiilor semantice** - sunt analizate interogări ale limbajului natural pentru importanța semantică a cuvintelor şi sunt extrase frazele relevante din toate documentele bazei de text. În plus, este format un subarbore de concepte referitoare la interogare, care facilitează o căutare simplă şi rafinată. O rețea semantică este o mulțime a celor mai importante concepte din text şi a relațiilor dintre aceste concepte cu ponderi în funcție de importanța lor relativă.

The size of the summary is controlled through the semantic weight threshold.

- **Subject-focused text exploration** - user-specified dictionaries of excluded and included words allow the investigation to focus on a chosen subject.
- **Efficient navigation through a textbase** - the knowledge base can be navigated with hyperlinks from concepts in the Semantic Network to sentences in the documents that contain the considered combination of concepts. Individual sentences are in turn hyperlinked to those places in original texts where they have been encountered.
- **Explication of the text theme structure** - a tree-like topic structure representing the semantics of the investigated texts is automatically developed. The more important subjects are placed closer to the root of a tree.
- **Clustering of texts** - breaking links representing weak relations in the original Semantic Network enables clustering of the textbase.
- **Semantic information retrieval** - natural language queries are analyzed for semantically important words and all relevant sentences from the textbase documents are retrieved. In addition, a subtree of concepts related to the query is formed, which facilitates a simple search refinement. A Semantic Network is a set of the most important concepts from the text and the relations between these concepts weighted by their relative importance.

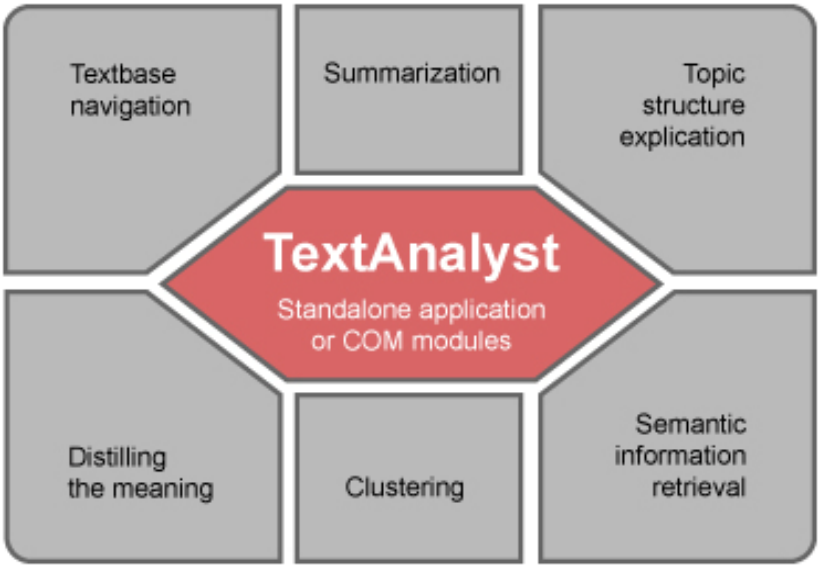

Figura 4. Sarcinile realizate de Text Analyst
Figure 4. Text Analyst tasks

Printre utilizatorii existenţi ai TextAnalyst se numără birourile guvernamentale, firme de consultanţă şi de drept, centre medicale, organizaţii ştiinţifice, editorii de carte electronică, centre de asistenţă pentru clienţi, instituţiile politice, şi chiar studenţii.

Existing users of TextAnalyst include government offices, consulting and law firms, medical centers, scientific organizations, electronic book publishers, customer support centers, political institutions, and even college students.

REFERINŢE

[1] J. Froelich, S. Ananyan, D.L. Olson, *"The Use of Text Mining to Analize Public Input", 2002*
[2] S. Ananyan, M. Kiselev, *Automated Analysis of Unstructured Texts, 2004*
[3] http://www.megaputer.com
[4] http://www.tlab.itAscultaţi

REFERENCES

[1] J. Froelich, S. Ananyan, D.L. Olson, *"The Use of Text Mining to Analize Public Input", 2002*
[2] S. Ananyan, M. Kiselev, *Automated Analysis of Unstructured Texts, 2004*
[3] http://www.megaputer.com
[4] http://www.tlab.it